Whose feedback?

Whose feedback? A multilevel analysis of student completion of end-of-term teaching evaluations

Leah P. Macfadyen*

Faculty of Arts, The University of British Columbia, Buchanan C110, 1866 Main Mall, Vancouver, BC Canada V6T 1Z1. leah.macfadyen@ubc.ca

Shane Dawson

University of South Australia, Learning and Teaching Unit, Level 1 David Pank Building, Adelaide, SA, Australia, 5001. Shane.dawson@unisa.edu.au

Stewart Prest

Department of Political Science, The University of British Columbia, C425, 1866 Main Mall, Vancouver, BC, Canada V6T 1Z1. stewartprest@gmail.com

Dragan Gašević

Schools of Education and Informatics, University of Edinburgh, Old Moray House, Holyrood Road, Edinburgh, EH8 8AQ, Scotland. dgasevic@acm.org

*Corresponding author

Author Biographical Notes

Leah P. Macfadyen is Program Director for Evaluation and Learning Analytics in the Faculty of Arts at The University of British Columbia, Canada. Her current research and collaborations are focussed on helping educational institutions make meaningful use of teaching and learning data through predictive modelling, data visualization and policy and strategy development.

Shane Dawson is Associate Professor, and Director of the Learning and Teaching Unit at the University of South Australia. Shane's research merges learning analytics and social network disciplines.

Stewart Prest is a PhD candidate in the department of Political Science at The University of British Columbia, specializing in the study of both civil and international peace and conflict. He has an extensive methodological background, including applied statistical research.

Dragan Gašević is Professor and Chair of Learning Analytics and Informatics in the Schools of Education and of Informatics at the University of Edinburgh, Scotland. His research seeks to understand and enhance social and self-regulatory aspects of learning with technology and learning analytics.

No potential conflict of interest is reported by the authors.

Whose feedback? A multilevel analysis of student completion of end-of-term teaching evaluations

Student evaluations of teaching (SETs) are now common practice across higher education, with the results used for both course improvement and quality assurance purposes. While much research has examined the validity of SETs for measuring teaching quality, few studies have investigated the factors that influence student participation in the SET process. This study aimed to address this deficit, through the analysis of an SET respondent pool at a large Canadian research intensive university. The findings were largely consistent with available research (showing influence of student gender, age, specialization area and final grade on SET completion). However, the study also identified additional influential course-specific factors such as term of study, course year level and course type as statistically significant factors influencing student response/non-response. Collectively, such findings point to substantively significant patterns of bias in the characteristics of the respondent pool. Further research is needed to specify and quantify the impact (if any) on SET scores. We conclude, however, by recommending that such bias does not invalidate SET implementation, but instead should be embraced and reported within standard institutional practice, allowing better understanding of feedback received, and driving future efforts at recruiting student respondents.

Keywords: SET; student evaluation of teaching; course evaluation; response rate; response bias; multilevel analysis

Introduction

Few practices in educational settings evoke emotional debate as rapidly as student evaluation of teaching (SET). While most educators acknowledge the value and importance of creating

opportunity for student feedback, many question the legitimacy of such forms of assessment when used for performance management and quality assurance (Stowell, Addison, and Smith, 2012). Opposition to the use of SETs for managerial purposes stems from perceived biases of different 'kinds' of student (Centra and Gaubatz, 2000), falling response rates (Adams and Umbach, 2012) and a perception that students may lack the maturity and expertise to provide informed and accurate feedback relating to teaching practice (Bedgood and Donovan, 2012; Clayson, 2009). The potential educational benefits derived from SETs are therefore often overshadowed by a powerful and pervasive belief among educators that they merely report on teacher popularity, rather than offering any rigorous measure of instructional effectiveness (Aleamoni, 1987; Feldman, 2007).

At the same time, the higher education sector in many countries has shifted towards a more business-oriented model of operation (e.g. Marginson and Considine, 2000; Mazzarol, Soutar, and Seng, 2003) in recent decades. As part of this transformation, the demonstration of institutional 'quality' has increasingly become a routine part of academic life. For this reason, it is unlikely that the current and common usage of SETs as a tool for measuring quality of teaching will diminish (Blackmore, 2009). Given the centrality of SETs in contemporary academic management and quality assurance processes, it is therefore increasingly important to ensure their validity, and to monitor and report on any potential response biases that may unduly influence SET results (for example, over- or under-representation of sub-populations of students as defined by gender, grade-point achievement or age grouping). Numerous researchers have interrogated institutional SETs by exploring response rates, or the characteristics or validity of the questionnaire employed (Marsh, 2007; Spooren, Brockx, and Mortelmans, 2013; Wachtel 1998), concluding generally that any bias is contextual and reflective of the institution itself, its ethos and culture.

In the present study we have employed a multi-level model of statistical analysis to investigate who did and did not respond to an institutional SET, with the goal of determining whether response bias may be influencing course design decisions or assessment of teaching performance to a practically significant degree.

Student evaluations of teaching

Student evaluations of teaching (SET) are common practice across the higher education sector. While SETs were initially introduced as part of an effort to improve teaching practice in the 1920s, the instrument has continued to evolve and usage has been extended into performance management practice (Galbraith, Merrill, and Kline, 2012). Marsh (2007) outlines five key applications of SET:

- Provision of diagnostic feedback for teachers
- Measurement of teaching effectiveness
- Provision of information for students regarding future course selections
- Quality assurance
- Pedagogical research

Most frequently, SETs are implemented as both a means of assessing effectiveness of course design, and also as instruments for performance management of instructional staff (appointment, promotion, tenure and quality assurance). It is this intersection of pedagogical and managerial functions that has caused so much friction in the academy. This apparent dual role – SETs as both developmental process and managerial/QA tool – has catalyzed high levels of sustained research and debate (Blackmore, 2009; Clayson, 2009). Almost three decades ago, Marsh (1987) noted that SETs are probably "the most thoroughly studied of all

forms of personal evaluation" (p.369). Research interrogating the validity and application of SET has continued unabated, and has produced a voluminous and contentious literature.

The possible connections between grading practices and SET scores has produced a multitude of what Aleamoni (1987, 1999) and Feldman (2007) call half-truths and myths in the academy. Commonly, critics assert that educators with a reputation for easy grading and a light course load will receive more favourable SET scores than their less lenient colleagues (Greenwald and Gillmore, 1997). This hypothesis has been extensively debated in the research literature. For example, McPherson and colleagues (2006; 2007; 2009) demonstrated a significant positive relationship exists between student course grades and SET scores. In a similar study investigating the factors affecting SET scores, Brockx et al (2011) also identified a significant positive relationship between grades and evaluation scores, but these authors contend that this correlation is underpinned not by grading leniency but by effective teaching practice. As Feldman (2007) explains, "students who learn more earn higher grades and thus legitimately give higher evaluations" (p.99). In other words, effective teaching facilitates student learning and this is reflected in higher levels of academic performance. In truth, any identified relationship between student grades and SET scores can be interpreted from multiple perspectives to either deny or confirm response bias, and herein lies the problem. While such studies can clearly and effectively identify if a significant relationship exists, it is much more difficult to design an empirical study that can confirm or deny causality.

Some findings are now well documented, however, and SETs are generally considered to be multi-dimensional valid indicators of teaching performance and effective for informing and improving teaching practice and course design (Marsh, 2007). The sheer volume of studies confirming the validity of student evaluations of teaching prompted Marsh (1987) to suggest that future SET research should focus on methodology, teaching context

and the characteristics that could negatively impact validity. One such area is response rates (Spooren and Van Loon, 2012). This is particularly topical as contemporary higher education institutions shift their evaluation practices from paper-based to online submissions (Anderson, Cain, and Bird, 2005).

Student response rates

With many universities now opting for online SETs, an associated decrease in student response rates has been reported (Stowell et al., 2012). A drop in response rates is understandably of grave concern for the everyday practitioner, especially when (re)appointment, promotion and tenure processes lean heavily on these forms of feedback and evaluation. To test for differences in the submission process, Stowell et al. (2012) compared response rates, SET scores and number of written responses to open-ended questions for online and paper-based SETs. These authors reported that although online response rates had fallen, there was no difference in overall average instructor ratings or written comments, confirming the findings of many earlier studies (Avery et al., 2006; Dommeyer et al., 2004; Layne, De Cristoforo, and McGinty, 1999) which had previously demonstrated no significant difference in mean instructor ratings despite lower response rates using online submissions. It is important to note, however, that there is an obvious minimum threshold for response rates beyond which the validity of evaluation scores is affected by the non-representativeness of the respondent sample (Dillman et al., 2002). In spite of the findings reported above, declining SET responses rates therefore continue to fuel academic mistrust and cynicism, allowing critics to call into question the validity of SETs.

The studies reported here highlight three important factors that should guide on-going SET research. First, the medium (online or paper-based) does not appear to unduly influence student ratings. Second, there is a need for research that can identify practices and processes

that can help address declining student response rates. Third, it is important to continue to monitor for possible response bias (Adams and Umbach, 2012), especially given the widespread application of these forms of teacher assessment in performance management. In this context, the online medium offers an advantage over older paper-based systems by capturing information about which students did or did not participate in the evaluation process. Investigating the sub-populations of non-responders and responders has the potential to reveal any possible bias that may affect instructor ratings. While SET research extends back many decades, few studies have examined the student or course characteristics that influence the decision to complete an SET.

The current study

The aim of this study is to provide further insight into factors that influence student response or non-response to SETs. Few studies have investigated the impact of course 'type' (e.g. lecture, independent study, experiential learning or group work), or the timing (point in the academic year) of SET implementation on student response decisions. To address this deficit we cross-tabulated data and performed simple logistic regression and multi-level linear modelling analyses to test the effect of these factors on student response/ non-response decisions. We also investigated the influence of student-specific factors (academic performance, gender, degree program, subject area specialization) and other course-specific factors (class size, course year level, salience with student specialization) on SET completion rates at the institution under study. By using a multi-level model design we aimed to identify any clustering effects and quantify the variation that may exist at the level of individual evaluation, individual course or aggregated group (e.g. school, program or cohort).

Methodology

A sample of end-of-term SET completion/non-completion data was collected from all courses offered in the Faculty of Arts at a large research-intensive Canadian university. The sample included selected data items for all students enrolled in at least one undergraduate course in the academic year 2009-2010, and for all course enrollments by these students. All SETs had been administered via an online evaluation system. Students within the Faculty of Arts were invited to complete one SET per enrolled course at the end of each teaching term in the academic year. Because each student may have been enrolled in multiple courses within the same time frame, individuals may have had the opportunity to complete multiple SETs. From a possible 94,161 course enrollments by 21,534 unique students, a total of 46,774 end-of-term SETs were completed, providing an overall average completion rate of 49.7%. The students in the sample were enrolled in the following degree program areas: Arts (N=10,426), Medical/Paramedical (N=32), Science (N=8,108), Education (N=24), Business (N= 1,862), and Fine Arts (N=446). Additional descriptive statistics are given in Table 1.

[Table 1 near here] □

Variables

For the purposes of this study the dependent variable was dichotomous (SET completion vs SET non-completion, for a given course SET). It is important to note that the institution's commitment to students on data privacy prohibits any access to data that link student identity to the SET scores or comments they submit. Available data does, however, allow us to link details of student identity and course enrollment record with their *completion* of each available SET survey. Available data for respondents/non-respondents includes: student age, student gender (coded 1 for male and 0 for female), final grade per course enrollment

(specified as both percentage and letter), student degree program, and student area of specialization (for example, a Major or Minor). Data relating to each course was also captured and reported, including course type (individual study, experiential, lecture-based, and small group), the total number of students enrolled within each course, and term in which a course was offered (Term 1, Term 2 or Term 1&2). A course term was therefore assigned one of three values: 0 for a two-term course, 1 for a Term 1 course, and 2 for a Term 2 course. The "term 1 or 2" variable captures the effect of the evaluation taking place in the first or second term as per the course schedule. The dummy variable "two-term course" (0) captures the influence of two term courses on student response or non-response.

Statistical analyses

The aim of the study was to test the association of the dependent (binary) variable with variables relating to both student and course. While cross-tabulations (see section 3.1) can reveal simple correlation patterns, they are unhelpful in situations where multiple influential factors at play. In such situations, a commonly adopted approach involves the use of a logit regression (Hosmer, Lemeshow, and Sturdivant, 2013), and we report findings from a simple logistic regression analysis of the data in section 3.3.

However, even logistic regression does not adequately account for the cross-classified hierarchical structure of the data analysed for this study (Hox, 1994; Hox and Kreft, 1994). A logit regression treats each observation (that is, an SET completed by a student) as unrelated to any other. However, given the nature of our sample set and context of SET implementation at this institutions – learners enrolled in different degree programs, who may have declared particular subject area specializations, grouped in course sections, being invited to complete SETs simultaneously implemented at end of term, and multiple course enrollments per student – we might hypothesize that the data could or should be grouped in meaningful ways

to provide richer insight into the rationale for response or non-response. The factors that influence student completion of SETs may be a function of a student's particular experience in a class, of a student's general disposition, or else a more general result of the properties of the section or class. The adoption of a multilevel model design allows us to capture such clustering effects, and to assign the level of variation that occurs at the level of evaluation, individual, and group. Recognizing the limitations of standard regression models, this method of multilevel analysis has also been proposed in educational research such as computer-supported collaborative learning (Cress, 2008; De Wever et al., 2007; Friend Wise, Saghafian, and Padmanabhan, 2012) and student evaluation of teaching (Adams and Umbach, 2012).

To determine the most meaningful grouping for analysis of our data, we ran an "empty model", often referred to as a variance-components model, which incorporated a number of different specifications (course section, course, student, course type, and degree type) in order to determine the relative variance occurring at different levels of analysis. This in turn allowed us to identify those levels that are most relevant in explaining the observed patterns of SET submission as well as those that lack significant explanatory power. The latter were then excluded from further analysis. In our analysis, we calculated two measures as shown in Table 2. First, ρ provides the residual intra-class correlation of the latent responses of a given model; this is a measure of the relative variance between and within groups. The larger ρ is, the greater the proportion of observed variance that occurs at the level of the group rather than the individual. The remaining information reported in Table 2 includes the estimate of the between-group standard deviation of the random intercepts of groups $\sqrt{\hat{\psi}}$, and the likelihood-ratio test of the hypothesis that ρ=0.

[Table 2 near here]

Most significantly, we found that nesting observations 'by student' captures far more of the total variance than any other hierarchical structure. While the likelihood ratio tests confirm that in all six cases ρ is statistically different than zero, the ρ of 0.769 obtained when nesting data by student far exceeds all other possibilities explored (Table 2). This can be interpreted to mean that 76.9% variance of the outcome variable was explained by the differences at the level of student. No other model achieved a ρ >=0.05 (i.e., less than 0.5% of the variance in the outcome). Based on this observation, we used 'by student' as the grouping variable in our multilevel analysis.

Results

Cross tabulation

Table 3 summarizes uncontrolled SET completion rates by grade and course year level. Each cell in the table provides the number of observations and the proportion of positive responses (mean completion rate) for that category (letter grade). The results suggest that response rates tend to be higher among students in years one and four, with students in year two, and to a lesser extent in year three, responding at a lower rate. Additionally, there is a clear and remarkably linear correlation between grades received and the likelihood that students respond. Completion rate increases as grade point does.

[Table 3 near here] □

Disciplinary salience

In their study of SET response rates Adams and Umbach (2012) found that *disciplinary salience* – the degree to which a particular course is aligned with an individual student's chosen disciplinary major – is an important predictor of student completion of an evaluation

survey. For example, a student with a major in history would be more likely to complete a SET for a history course than an alternate course outside of this primary disciplinary area. We investigated disciplinary salience in the study sample by reviewing the SET completion rates of students who had declared a subject specialization (Major, Minor or Honours), and only for those subject specializations where N(enrollments)>15. This reduced the sample to a set of 5,706 unique students with 36,673 course enrollments (and thus SET invitations). For each specialization group, we calculated overall SET response rate for courses within the specialization area, as well as overall SET response rates for courses completed in all other subject areas.

As shown in Figure 1, our findings tend to support those of Adams and Umbach (2012). In fifteen of eighteen student specialization areas, students completed SETs for courses in their specialization area with a response rate 1-22% higher than their completion of SETs for courses in other subject areas.

[Figure 1 near here]

Simple regression analysis of SET completion data

Table 4 shows our base logit regression model (and accompanying variations) of the binary dependent variable completed on the variables student age, course year level, course year level squared, gender (coded 1 for male and 0 for female), term, a dummy for two-term course, percent grade achieved, and class size (ln(enrollment), the natural log of enrollment). The squared course year level variable is included to capture the non-linear nature of the relationship highlighted in Table 3. From this analysis, all indicators are found to be significant at the 1% level. Age and final grade both have a positive effect on the likelihood that a student will complete an SET, while term and class size have a negative effect. Course year level appears to have a non-linear relationship, whereby completion rates initially

decrease as year level increases, but begin to increase at higher levels.

To modulate the initial findings, we introduced variables such as course type and student degree program. As these variables do not lend themselves to ordinal analysis, we incorporate each into a separate model, also included in Table 4. Initial results from model 1 are robust to these alternative model specifications, with both coefficients and standard errors remaining relatively constant. Some interesting findings emerge from these additional results. Most significantly, perhaps, students in lecture-based courses are more likely to submit evaluations than any other course type. Furthermore, this finding is significant at the 5% level (or better) for all course types. A second notable finding is that students in Arts degree programs are less likely to submit responses than any other student degree type, save Fine Arts. This finding is statistically significant at the 1% level for Science and Medical/Paramedical students and at the 5% level for Business and Fine Arts students. Education students are statistically indistinguishable from their Arts counterparts.[1]

[Table 4 near here]

Using the simulations provided by Stata's *Clarify* program (Tomz, Wittenberg and King, 2003), we can further quantify the substantive effects of these results. Table 5 provides a range of probabilities that result from a given change in a specific variable, holding all other variables at their median values. The column "mean" provides the average change in probability for a given student for a given change in one explanatory variable, holding other variables at the median. The standard error term indicates the relative statistical significance

[1] There is, in principle, a risk of a false discovery in these findings, though the high degree of significance consistently reported across models tends to render the possibility remote. We would be much more concerned if, in the multiplicity of tests, we had found only one, or a small set of findings that achieved statistical significance. More generally, the fact that we include results from a multilevel modelling approach further mitigates the risk of false discovery due to familywise error in our reported results, insofar as the grouping of units constitutes a form of "partial pooling" tending to make estimates more appropriately conservative, but not excessively so as with traditional methods of control such as Bonferonni correction. See Gelman, Hill, and Yajima (2012) for more information. That said, continuing research on the subject matter must remain cognizant of the risk of familywise error in reported findings involving multiple hypothesis tests, and use some strategy to manage accordingly.

of the result. The final two columns provide the 95% confidence interval, which is the range within which the true value of the coefficient would be found, 19 times out of 20. If the interval includes 0, the result may be considered statistically insignificant at the 5% level.

The first row – "probability at the median" – gives the mean probability of submission, holding all controlled variables at their median values, on the basis of model 1, our base model. Thus, for the median student in the base model, the probability of submission is 0.54. The other lines all give the effect of a specific change in values, holding other variables at their median scores. For example, the probability of submission for a male student is 0.075 less than for a female student. Students in full term courses are 0.17 less likely to complete an evaluation than students in term 1, while students in term 2 are 0.09 less likely to do so than their term 1 counterparts. The probability of a student in a first year course completing an evaluation is 0.098 higher than a student in a fourth year course, and 0.106 higher than a student in a third year course. Students in sections ranking in the 10th percentile of class size – 24 students – have a 0.085 greater probability completing an evaluation than those in sections in the 90th percentile, with enrollments of 245 students. Finally, students scoring in the 10th percentile in terms of grade – who received 57%, or a D – were 24% less likely to complete an evaluation than their counterparts in the 90th percentile, who received 86%, or an A.

[Table 5 near here]

Examining selected results from models 2 and 3, a change from a lecture-based course to an experiential course results in a 0.06 decrease in the relative probability of submission. Conversely, a change from lecture-based to small group results in a less than 1% change in the probability; this result (as with the coefficient in Table 4 above) is not significant. A Science student is 0.097 more likely to submit an evaluation than an Arts

student, holding other values at the median, while a Medical/Paramedical student is 0.19 more likely.

Multi-level analysis

Table 6 presents the base hierarchical model, a logit regression clustered by student that includes coefficients for age, course year level, gender, two-term courses, percent grade, and class size, along with student mean of class size, grade and term (Inclusion of these means allow us to isolate the so-called "between" and "within" effects of each variable and identify the effects of covariates that vary across different observations for a given student, as well as across different students). The coefficients in Table 6 represent the change in log-odds due to a unit change in a given variable while holding other variables constant at the mean, but are difficult to interpret directly. We therefore also include in Table 6 the odds ratio for the covariates. These may be interpreted as the likelihood of a positive outcome (i.e. SET completed) divided by the likelihood of a negative outcome. For example, using the basic model in Table 6, the odds ratio for student age is calculated to be 1.061. Thus, holding all other values at (any) fixed values, for each additional year in age, the odds of a student completing an SET is 1.061, or 6.1% higher than a student one year younger.

[Table 6 near here]

To further aid interpretation, Table 7 reports on the effects of discrete changes in the values of selected covariates, while holding other variables constant. Again, the results are expressed in odds ratios, which assist the interpretation over the raw coefficients.

[Table 7 near here]

Academic performance (grade achievement) was observed to have an additional significant effect on response/ non-response (Table 7). A change from the 10^{th} to the 90^{th} percentile in course grade, (or from 57 to 86% percent grade) increases the odds of response

by 1.65, or 65%. That is, a student is 65% more likely to respond in courses that they do well in than those that they do not. Hence, if a given student has 0.33 probability of responding in a course in which they receive 57%, that student would have a 0.54 probability of responding in a course in which they received 86% percent grade. This effect is substantively greater between students. The odds of response of a student with an 86% average percent grade are 7.6 times greater than that of a student with a 57% average percent grade. This is a significant and notably large effect. It is useful to convert the results to an absolute probability, using the formula probability=odds/(1+odds). Having done so, the probability of response by a student at the 90^{th} percentile is 0.88 greater than for a student at the 10^{th} percentile. Finally, the individual level effect of class size obtains as greater than the population average effect. That is, the effect of increasing class size is greater on an individual student from one course to the next, than it is for the average class size experienced from one student to the next.

Finally, in Table 8 we report the results obtained when variables for the specific type of course are included. "Lecture-type courses" constitutes the base category, and all other results are evaluated as deviations from the odds of response for a student in a lecture based course. Note that we did not include the covariates for the mean of each dummy; thus we cannot say whether the effect is greater between or across students. However, the effects reported are, as above, subject-specific effects, rather than the population-specific results as would be the case had we adopted the standard logit regression for our analysis. The reported odds of a given student responding, when changing from a lecture environment to an experiential one is 0.696; conversely, the odds of response for a student moving from an experiential environment to a lecture environment are in fact 1.42 times, or 42% greater. The effect for individual course type is also negative and is actually substantively larger, but it is significant only at the 10% level. The effect for small group course is not statistically significant.

[Table 8 near here]

Discussion

In this study we sought to examine the impact of a range of factors specific to the selected institutional context on student SET completion rates, and to test whether other factors reported in the small volume of literature on student completion of SETs are also relevant.

Our findings confirm that a range of student-specific factors influence the likelihood that a student will complete an SET. Prior studies have noted that age (Spooren and Van Loon, 2012), gender and disciplinary salience (Adams and Umbach, 2012) are potential factors associated with response bias. In the current sample, the odds ratio of SET completion by student gender was determined to be 0.580: that is, other things being equal, the odds of a male student submitting an SET are 0.58 times that of a female student. We also found that for a change from the 10^{th} to the 90^{th} percentile in age (which in this sample represents a shift from age 19.9 to age 25.4), the odds ratio is 1.39 (Table 6). Moreover, our investigation of disciplinary salience for this sample showed that students are more likely to complete SETs for courses coherent with their declared degree specialization area (for example, a declared Major, Minor or Honours subject) (Figure 1). In other words, and as reported by others, older students, female students and students enrolled in courses relevant to their study specialization are over-represented in the respondent pool. The implication is that a particular decision to submit an SET is more strongly influenced by individual-level characteristics, rather than by factors relating to student degree program, type of course or the course itself. This is borne out by our variance-components analysis model which confirmed that nesting variables at the level of 'the student' offers the greatest explanatory power for the variance in response rates we observe.

The factor most commonly argued to influence student SET response/ non-response is academic performance (as represented by final grade achieved in a course). By simple cross-tabulation, we found that there is a clear and remarkably linear correlation between final letter grade achieved in a course and the likelihood that a student completes the associated SET (Table 3), consistent with the findings of Adams and Umbach (2012) and Spooren and Van Loon (2012). This positive correlation persists even when differences in student age, gender, degree program and course year level, type and term are controlled for (Tables 4 and 5). Interestingly, while this effect holds true between students, and also within an individual student's multiple course SETs, our multi-level analysis demonstrated that the effect is greater 'within' a student's record. That is, an individual student is more likely to complete SETs for courses in which they ultimately achieve a higher final grade. Because learners at the institution under study must make the SET completion/non-completion decision *before* completing final assessments or receiving final grades, we suggest that in this context SET completion (and scores awarded) are not simple pleasure/displeasure responses by students to grade 'reward'/'punishment' by instructors. Rather, we propose that final grade can be considered a proxy for a student's overall learning experience, which in turn may influence SET completion. Extending this logic, Spooren and Van Loon (2012) have argued that the relationship between final grade and SET completion may in part explain the observed positive correlation between final grades and instructor scores that has been identified in multiple studies (e.g. Brockx et al., 2011; McPherson, 2006; McPherson and Jewell, 2007; McPherson et al., 2009). That is to say, SET completion itself may be an indicator of 'student satisfaction', and any bias in scores introduced may be skewed in favour of positive scores. At a minimum, it is clear that the observed relationship between grades, SET completion and SET scores is complex and requires further research into areas such as student decision making processes and motivations.

The relationship between other aspects of a course experience and a student's decision to complete an SET or not is also less than straightforward. We found that class size is moderately and negatively correlated with SET completion, and similarly to the grade effect, this relationship holds true between and within students. In addition, students in courses coded as 'individual study' are less likely to complete an SET – a finding that supporters of social constructivist theories of learning might interpret as supporting the premise that 'good learning' is social and thus requires peers. This is confounded, however, by our finding that students in 'traditional lecture' courses – commonly argued to be less engaging (Marsh, 1987) – are 42% more likely to complete an SET. In the current context, this may be relieving for educators, given that the vast majority of enrollments (94% of the current sample) are in courses coded as 'lecture-based', but this finding does not illuminate the nature of the connection between course type, the learning experience, and student decisions around SET completion. Are lecture-based courses simply a more familiar learning environment for students and thus more likely to promote 'satisfaction' and SET completion? Or might 'group' forces, instructor communications to the group or peer communications simply facilitate higher rates of SET completion (than for individual, small group or experiential courses)?

Some of our observations might be interpreted as indicative of 'evaluation fatigue'. In the short-term, it appears that by the end of a second term of study in an academic year, students are less likely to complete SETs. One interesting result that emerges from the inclusion of both Term and Term mean variables in our random effects model is that the Term effect is different within and across students. That is, for a given student the effect of moving from Term 1 to Term 2 *decreases* the probability of responding. These results suggest that the drop in response rates from Term 1 to Term 2 does not represent something intrinsically different about Term 2 courses, but rather that the act of completing Term 2

courses after Term 1 courses reduces the likelihood of response. Further research is required to determine the extent of this effect through interaction of course year level and term variables, or inclusion of a dummy variable for students who are enrolled in Term 2 courses only.

The multi-level analysis also indicates that students in first year courses complete SETs more frequently, however, this response rate drops as students progress through their degree program. once we control for factors such as student grade and class size (not shown here), the effect becomes more clearly negative and monotonic, with the biggest decline coming between years 1 and 2. Students in third year courses are in turn marginally less likely to respond than students in second year courses, while students in fourth year courses are marginally less likely to respond than those in third year courses. Overall, then, we have confirmed that a statistically significant degree of response bias exists in the current institutional sample, though without further investigation the effect of this bias on SET scores and thus on course design decisions or assessment of teaching performance remains unclear. Additional research is needed to further specify exact sources of bias, and to quantify their effects on evaluation.

Conclusions

What can we learn from such confirmation of response bias? What are the implications for evaluative practice in the institution, and the reliability of SET for both pedagogical and management uses?

First, our findings indicate that the fears of anxious and often angry academic staff who oppose evaluation are to some degree confirmed. Respondent pools do not fully represent the distribution of students in courses, and while the impact of this non-representativeness on SET scores has not been demonstrated (and may even skew scores

positively), such response bias is sufficient to fan the flames of suspicion. Greater efforts to improve recruitment of students for SET completion are warranted. Clearly, we have no ability to regulate innate propensities of individual students that may depend on age, gender or even study choices. And institutional and budgetary constraints may limit capacity to make significant changes to class size or range of course types available. But a cynic might point out that savvy time-constrained students, bombarded with survey requests throughout the academic year and throughout their programs of study, are likely to make rational decisions about whether to invest time in completing SETs based on the perceived level of benefits returned. In the context under study, the institution has adopted few formal strategies to report back to students the findings of evaluations, or to demonstrate any resultant action taken as a result of student evaluations of teaching. The need to close the feedback loop in the SET process is, however, increasingly evident: completion rates may improve if students perceive that feedback from SETs is reviewed and valued, and that it carries real import for modifying and improving their overall learning experience (Bennett and Nair, 2010; Nair, Adams, and Mertova, 2008). There is an obvious need for the institution represented in this study to better demonstrate the importance it places on feedback derived from the student body. Requiring or otherwise incentivizing SET completion also has the potential to improve SET completion rates.

Importantly, acknowledging the likely connection between demonstrating the benefit of SETs to learners and SET completion rates highlights the rarely-acknowledged reality that – in the context of voluntary SETs – the 'performance and management' usage of SETs is highly dependent on their real and demonstrated usage for diagnostic, educational and pedagogical purposes. Institutions remain dependent on SET output for quality assurance and performance management processes, even as SET completion rates decline and demonstrate response bias. Demonstrating to students that their feedback offers real benefits to themselves

has the potential to sustain this multi-purpose system of course and teaching evaluation and ensure that its output is valid and reliable.

In summary, this study demonstrates that a student's decision to complete a SET is not a random process. There are multiple course-, teacher- and student-specific factors that influence the decision to participate in the SET process. Here we propose that as part of good professional and institutional practice, any demonstrated bias in respondent pools should be reported and acknowledged. Making such data available and transparent, alongside institutional recognition of the complexities associated with these forms of evaluation, may serve to legitimize SETs within academic practice and culture. As long as SETs continue to play an important role as indicators of teaching quality, and as long as they are used to generate data in support of (re)appointment, promotion and tenure applications, it is critical that information regarding potential survey bias is included any presented reports.

Given the wide range of psychological, social, cultural, and pedagogical factors that can influence a student's decision to engage or not engage in the evaluation process, some response bias should not be surprising. We argue, however, that bias (or more correctly, the characteristics of the respondent pool) should also be embraced and incorporated into all discussions regarding teaching quality and course improvement. At present instructors receive course feedback and a statement of overall course response rates, implying (by omission) that the feedback obtained is representative of the entire course cohort. It is inferred from this that any subsequent course modifications are undertaken in the best interests of the course for any future student cohorts. We suggest that such inferences are misleading. Instead, inclusion in reports of analyses of the characteristics of responding and non-responding students may offer a valuable supplement to quantitative and qualitative feedback received. Such data would assist instructors with interpretation of their own evaluation results, better inform development of institutional strategies to recruit more

representative student feedback on SETs, and assist promotion and tenure committees in their decision making processes.

The key challenge for education systems lies in addressing how we can better motivate the student population to submit SETs. Clearly, these forms of evaluation play an important role in course and teaching improvement practice. By continuing to interrogate patterns of student response/non-response to SETs we can more effectively target under-represented student groups, promote to all students the benefits that are derived from teacher and course evaluations, and reassure academic staff of the value and reliability of evaluation data.

References

Adams, M. J .D., and P. D. Umbach. 2012. "Nonresponse and online student evaluations of teaching: Understanding the influence of salience, fatigue, and academic environments". *Research in Higher Education* 53 (5): 576-591.

Aleamoni, L. 1987. "Student rating myths versus research facts." *Journal of Personnel Evaluation in Education* 1 (1): 111-119.

Aleamoni, L. 1999. "Student rating myths versus research facts from 1924 to 1998." *Journal of Personnel Evaluation in Education* 13 (2): 153-166.

Anderson, H.M., J. Cain and E. Bird. 2005. "Online Student Course Evaluations: Review of Literature and a Pilot Study." *American Journal of Pharmaceutical Education* 69 (1): 34-43.

Avery, R. J., W. K. Bryant, A. Mathios, H. Kang, and D. Bell. 2006. "Electronic Course Evaluations: Does an Online Delivery System Influence Student Evaluations?" *The Journal of Economic Education* 37 (1): 21-37. doi: 10.3200/JECE.37.1.21-37

Bedgood, R. E., and J. D. Donovan. 2012. "University performance evaluations: What are we really measuring?" *Studies in Higher Education* 37 (7): 825-842.

Bennett, L., and C. S. Nair. 2010. "A recipe for effective participation rates for web-based surveys." *Assessment and Evaluation in Higher Education* 35 (4): 357-365.

Beran, T., and C. Violato. 2005. "Ratings of university teacher instruction: How much do student and course characteristics really matter?" *Assessment and Evaluation in Higher Education* 30 (6): 593-601.

Blackmore, J. 2009. "Academic pedagogies, quality logics and performative universities: Evaluating teaching and what students want." *Studies in Higher Education* 34 (8): 857-872.

Brockx, B., P. Spooren and D. Mortelmans. 2011. "Taking the grading leniency story to the edge. The influence of student, teacher, and course characteristics on student evaluations of teaching in higher education." *Educational Assessment, Evaluation and Accountability* 23 (4): 289-306.

Centra, J. A., and N. B. Gaubatz. 2000. "Is there gender bias in student evaluations of teaching?" *The Journal of Higher Education* 71 (1): 17-33.

Clayson, D. E. 2009. "Student Evaluations of Teaching: Are They Related to What Students Learn? A Meta-Analysis and Review of the Literature." *Journal of Marketing Education* 31 (1): 16-30.

Cress, U. S. 2008. "The need for considering multilevel analysis in CSCL research - An appeal for the use of more advanced statistical methods." *International Journal for Computer-Supported Collaborative Learning* 3 (1): 69-84.

Davies, M., J. Hirschberg, J. Lye, C. Johnston, C. and I. McDonald. 2007. "Systematic influences on teaching evaluations: The case for caution." *Australian Economic Papers* 46 (1): 18-38.

De Wever, B., H. Van Keer, T. Schellens, and M. Valcke. 2007. "Applying multilevel modelling to content analysis data: Methodological issues in the study of role assignment in asynchronous discussion groups." *Learning and Instruction* 17 (4): 436-447.

Dillman, D. A., J. L. Eltinge, R. M. Groves, and R. J. A. Little. 2002. "Survey non response in design, data collection, and analysis." In *Survey Nonresponse* edited by R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. Little, 3-26. New York: John Wiley and Sons.

Dommeyer, C. J., P. Baum, R. W. Hanna, and K. S. Chapman. 2004. "Gathering faculty teaching evaluations by in-class and online surveys: Their effects on response rates and evaluations." *Assessment and Evaluation in Higher Education* 29 (5): 611-623.

Feldman, K. A. 2007. "Identifying exemplary teachers and teaching: Evidence from student ratings." In *The Scholarship of Teaching and Learning in Higher Education: An*

evidence-based perspective edited by R. P. Perry and J. C. Smart, 93-143. Dordrecht: Springer.

Friend Wise, A., M. Saghafian, and P. Padmanabhan. 2012. "Towards more precise design guidance: specifying and testing the functions of assigned student roles in online discussions." *Educational Technology Research and Development* 60 (1): 55-82. doi: 10.1007/s11423-011-9212-7

Galbraith, C. S., G. B. Merrill, and D. M. Kline. 2012. "Are Student Evaluations of Teaching Effectiveness Valid for Measuring Student Learning Outcomes in Business Related Classes? A Neural Network and Bayesian Analyses." *Research in Higher Education,* 53 (3): 353-374. doi: 10.1007/s11162-011-9229-0

Gelman, A., J. Hill, and M. Yajima. 2012. "Why We (Usually) Don't Have to Worry about Multiple Comparisons." *Journal of Research on Educational Effectiveness.* 5: 189-211.

Greenwald, A. G., and G. M. Gillmore. 1997. "Grading leniency is a removable contaminant of student ratings." *American Psychologist* 52 (11): 1209-1217.

Hosmer, D. W., S. Lemeshow, and R. X. Sturdivant. 2013. "Introduction to the Logistic Regression Model". In *Applied Logistic Regression, Third Edition,* edited by Hosmer, D. W. , S. Lemeshow, and R. X. Sturdivant, 1-33. Hoboken, NJ, USA: John Wiley and Sons, Inc.

Hox, J. J. 1994. "Hierarchical Regression Models for Interviewer and Respondent Effects." *Sociological Methods and Research* 22 (3): 300-318. doi: 10.1177/0049124194022003002

Hox, J. J., and I. G. G. Kreft. 1994. "Multilevel Analysis Methods." *Sociological Methods and Research* 22 (3): 283-299. doi: 10.1177/0049124194022003001

Layne, B. H., J. R. De Cristoforo, and D. McGinty. 1999. "Electronic versus traditional student ratings of instruction." *Research in Higher Education* 40 (2): 221-232.

Marginson, S., and M. Considine. 2000. *The enterprise university : Power, governance, and reinvention in Australia*. New York: Cambridge University Press.

Marsh, H. W. 1987. "Students' evaluations of university teaching: research findings, methodological issues, and directions for future research." *International Journal of Educational Research* 11: 253-388.

Marsh, H. W. 2007. "Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases and Usefulness." In *The Scholarship of*

Teaching and Learning in Higher Education: An Evidence-Based Perspective edited by R. P. Perry and J. C. Smart, 319-383. Netherlands: Springer.

Mazzarol, T., G. N. Soutar, and M. S. Y. Seng. (2003). "The third wave: Future trends in international education." *International Journal of Educational Management* 17 (3): 90-99.

McKeachie, W. J. 1979. "Student ratings of faculty: A reprise." *Academe* 65 (6): 384-397.

McPherson, M. A. 2006. "Determinants of how students evaluate teachers." *Journal of Economic Education* 37 (1): 3-20.

McPherson, M. A., and R. T. Jewell. 2007. "Leveling the playing field: should student evaluation scores be adjusted?" *Social Science Quarterly* 88 (3): 868-881.

McPherson, M.A., R. T. Jewell, and M. Kim. 2009. "What determines student evaluation scores? A random effects analysis of undergraduate economics classes." *Eastern Economic Journal* 35 (1): 37-51.

Nair, C. S., P. Adams, and P. Mertova. 2008. "Student engagement: the key to improving survey response rates." *Quality in Higher Education* 14 (3): 225-232.

Spooren, P., B. Brockx, and D. Mortelmans. 2013. "On the Validity of Student Evaluation of Teaching The State of the Art." *Review of Educational Research* 83 (4): 598-642.

Spooren, P., and F. Van Loon. 2012. "Who Participates (not)? A Non-Response Analysis on Students' Evaluations of Teaching." *Procedia - Social and Behavioral Sciences* 69 (0): 990-996. doi: http://dx.doi.org/10.1016/j.sbspro.2012.12.025

Stowell, J. R., W. E. Addison, and J. L. Smith. 2012. "Comparison of online and classroom-based student evaluations of instruction." *Assessment and Evaluation in Higher Education* 37 (4): 465-473.

Tomz, M., J. Wittenberg, and G. King. 2003. "CLARIFY: Software for interpreting and presenting statistical results." *Journal of Statistical Software* 8 (1): 1-30.

Wachtel, H. K. 1998. "Student Evaluation of College Teaching Effectiveness: a brief review." *Assessment and Evaluation in Higher Education* 23 (2): 191-212. doi: http://dx.doi.org/10.1080/0260293980230207

Table 1. Descriptive statistics for sample

Variable		N	Mean	SD
Unique students		21,534	-	-
Female students		12,285	-	-
Enrollments		94,161	-	-
Female enrollments		57,804	-	-
Submitted evaluations per unique student			4.4	2.92
Percent grade			72.7	13.26
Student age			22.6	3.39
Enrollments by letter grade achieved				
	A+	4,841		
	A	9,434		
	A-	16,304		
	B+	14,360		
	B	13,919		
	B-	10,920		
	C+	7,804		
	C	5,460		
	C-	4,170		
	D	3,137		
	F	3,812		
Enrollments by course year level				
	1	25,292		
	2	21,622		
	3	25,378		
	4	21,869		
Enrollments by course term				
	1	43,475		
	2	42,147		
	1&2 (two-term)	8,539		
Enrollments by course type				
	Lecture-based	86,634		
	Experiential	1,374		
	Small group	3,790		
	Individual study	106		
Enrollments by student degree program type				
	Arts	66,617		
	Science	17,652		
	Education	84		
	Fine Arts	1,877		
	Medical/Paramedical	89		
	Business	5,585		

'Course type' categorization makes use of descriptive data collected and maintained by the university's enrollment services unit. 'Lecture-based' courses include those coded as Lecture-Discussion, *Lecture-Lab, Lecture-Seminar,* or *Lecture only*; 'Experiential' courses include those coded as *Field Trip, Lab, Practicum, Rehearsal,* or *Studio*; 'Small group' courses include those coded as *Seminar* or *Tutorial*; 'Individual study' courses include those coded as *Directed Studies, Essay/Research, Project, Thesis,* or *Project.*

Table 2. Variance components models for selected hierarchical structures

	By course section		**By student**		**By course**	
	Coefficient	Std. err.	Coefficient	Std. err.	Coefficient	Std. err.
Constant	0.093	0.012	0.133	0.027	0.080	0.014
$\sqrt{\hat{\psi}}$	0.384	0.011	3.308	0.040	0.326	0.013
ρ	0.043	0.002	0.769	0.004	0.031	0.002
Likelihood ratio test	χ^2=1692.4 Pr (ρ=0) < 0.001		χ^2=3.2*10^4 Pr (ρ=0) < 0.001		χ^2=1739.42 Pr (ρ=0) < 0.001	

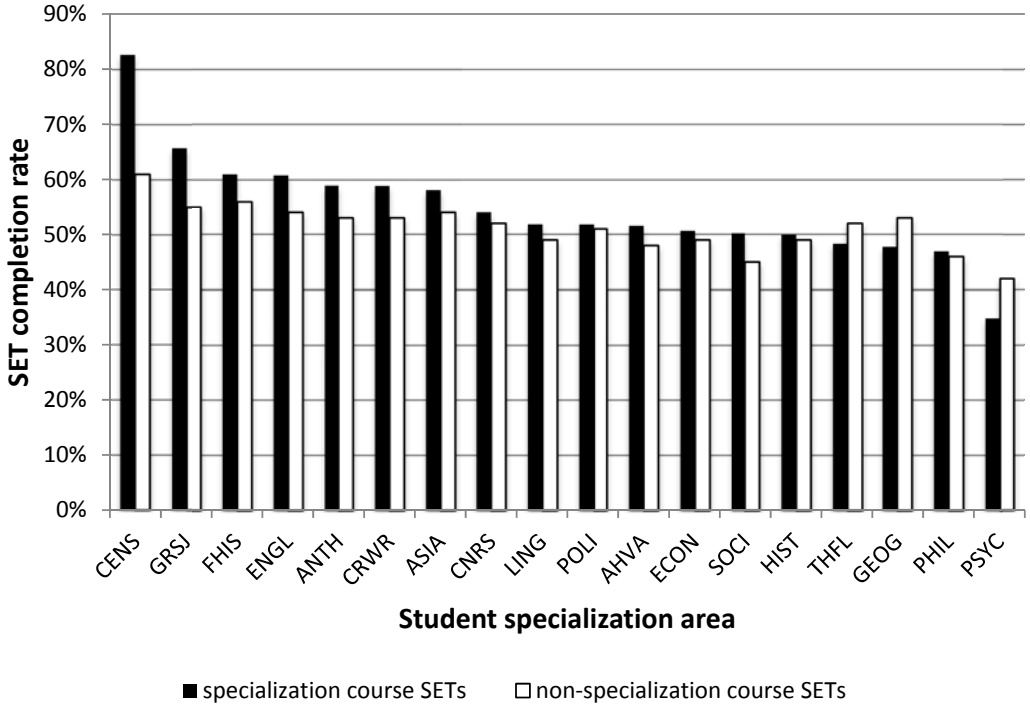
	By degree program		**By course type**	
	Coefficient	Std. err.	Coefficient	Std. err.
Constant	0.129	0.135	0.135	0.044
$\sqrt{\hat{\psi}}$	0.217	0.136	0.136	0.034
ρ	0.014	0.006	0.006	0.003
Likelihood ratio test	χ^2= 417.82 Pr (ρ=0) < 0.001		χ^2=106.07 Pr (ρ=0) < 0.001	

*Note: Reporting logistic random intercept models with completion as the dichotomous dependent variable.

Table 3. Cross-tabulation of observations and mean evaluation completion rate by letter grade and course year level (1-4)

Grade	**Year 1**	**Year 2**	**Year 3**	**Year 4**	**Total**
F	1,716 0.290	776 0.219	818 0.222	502 0.231	3,812 0.253
D	1,173 0.367	697 0.307	775 0.295	492 0.360	3,137 0.335
C-	1,501 0.396	972 0.343	989 0.323	708 0.356	4,170 0.359
C	1,829 0.425	1,336 0.389	1,332 0.375	963 0.428	5,460 0.404
C+	2,514 0.465	1,951 0.393	1,835 0.413	1,504 0.422	7,804 0.426
B-	3,224 0.518	2,656 0.448	2,875 0.455	2,165 0.465	10,920 0.474
B	3,789 0.532	3,384 0.478	3,704 0.488	3,041 0.472	13,918 0.494
B+	3,465 0.580	3,390 0.501	4,008 0.521	3,497 0.536	14,360 0.534
A-	3,302 0.613	3,553 0.543	4,810 0.557	4,637 0.565	16,302 0.568
A	1,761 0.633	1,948 0.583	2,798 0.599	2,927 0.611	9,434 0.606
A+	1,018 0.653	959 0.602	1,434 0.618	1,430 0.636	4,841 0.628
Total	**25,292** **0.513**	**21,622** **0.470**	**25,378** **0.490**	**21,866** **0.513**	**94,158** **0.497**

Figure 1. SET completion rates by student subject area specialization



CENS= European Studies; GRSJ=Gender Studies; FHIS= French, Hispanic & Italian; ENGL=English; ANTH=Anthropology; CRWR=Creative Writing; ASIA=Asian Studies; CNRS=Classics; LING=Linguistics; POLI=Political Science; AHVA=Fine Arts; ECON=Economics; SOCI=Sociology; HIST=History; THFL=Theatre & Film; GEOG=Geography; PHIL=Philosophy; PSYC=Psychology.

Table 4. Selected variations of basic logistic model

	Model 1: Base Model		**Model 2: Course type**		**Model 3: Degree type**	
Variable	Co-eff.	Std. err.	Co-eff.	Std. err.	Co-eff.	Std. err.
Student age	0.030***	0.003	0.031***	0.003	0.034***	0.003
Course year level	-0.556***	0.034	-0.557***	0.034	-0.515***	0.035
(Course year level)^2	0.084***	0.007	0.084***	0.007	0.077***	0.007
Gender (male=1)	-0.302***	0.014	-0.303***	0.014	-0.335***	0.014
Semester (1 or 2)	-0.355***	0.014	-0.355***	0.014	-0.352***	0.014
Two-semester course	-0.681***	0.032	-0.681***	0.032	-0.631***	0.032
Percent grade	0.029***	0.001	0.029***	0.001	0.028***	0.001
Class size (ln(enrollment))	-0.148***	0.008	-0.157***	0.008	-0.177***	0.008
Individual study	—	—	-0.426**	0.198	—	—
Experiential course	—	—	-0.258***	0.057	—	—
Small group course	—	—	-0.028	0.036	—	—
Medical/Paramedical degree	—	—	—	—	0.865***	0.240
Science degree	—	—	—	—	0.410***	0.018
Education degree	—	—	—	—	0.256	0.236
Business degree	—	—	—	—	0.065**	0.029
Fine arts / design degree	—	—	—	—	-0.114**	0.050
Constant	-0.772***	0.088	-0.741***	0.088	-0.793***	0.089
N	94158		94158		91904	
Log likelihood	-62688.2		-62675.7		-60913.1	
Pseudo R^2	0.040		0.040		0.044	
Coefficients marked with (***) are significant at the 1% level, with (**) are significant at the 5% level, and (*) at the 10%. All standard errors are White robust to account for heterogeneity.						
Notes:			Base category is Lecture course type		Base category is Arts degree.	

Table 5. Effects of discrete value changes in the probability of evaluation completion

Quantity of Interest	**Mean**	**Std. Err.**	**[95% Conf. Interval]**	
Model 1				
Probability at the median	0.544	0.003	0.537	0.551
Change from female to male	-0.075	0.003	-0.082	-0.069
Change from fullterm=0 to 1	-0.167	0.008	-0.182	-0.152
Change from term 1 to term 2	-0.088	0.004	-0.096	-0.081
Change from 10th-90th percentile in age	0.042	0.004	0.034	0.048
Change from course year 1 to year 4	-0.098	0.005	-0.108	-0.087
Change from course year 1 to year 3	-0.106	0.005	-0.115	-0.097
Change from 10-90th percentile in class size	-0.085	0.004	-0.094	-0.076
Change from 10-90th percentile in student grade	0.208	0.004	0.200	0.215
Model 2				
Change from lecture to experiential	-0.064	0.014	-0.092	-0.038
Change from lecture to small group	-0.007	0.009	-0.024	0.012
Model 3				
Change from Arts to Science	0.097	0.004	0.089	0.105
Change from Arts to Medical/Paramedical	0.188	0.047	0.089	0.271

Table 6. Random effects models for evaluation completion data – base model

Variable	**Co-efficient**	**Std. err.**	**Odds ratio**	**95% confidence interval**	
Fixed part					
Student age	0.059***	0.009	1.061	1.043	1.080
Course year level	-0.458***	0.029	0.632	0.598	0.669
Gender (male=1)	-0.544***	0.056	0.580	0.520	0.648
Term (1 or 2)	-0.906***	0.023	0.404	0.386	0.423
Mean of term by student	0.379***	0.076	1.462	1.260	1.696
Two-term course	-1.681***	0.052	0.186	0.168	0.206
Percent grade	0.017***	0.001	1.017	1.015	1.020
Mean of grade by student	0.070***	0.003	1.073	1.066	1.079
Class size (ln(enrollment))	-0.280***	0.014	0.756	0.735	0.778
Mean of class size by student	-0.141***	0.044	0.869	0.797	0.947
Constant	-3.528***	0.367			
Random part			*Other statistics*		
$\sqrt{\hat{\psi}}$	3.30		N	94158	
P	0.767		Groups	21533	
Log likelihood	-47357.4		Obs/group	4.4	

Notes: Base category is the lecture-based course type.

Table 7. Base model changes in odds ratios from changes in selected covariates

Variable	**Change**	**Co-eff.**	**Std. Err.**	**Odds Ratio**	**95% confidence interval**	
Age	From 10th to 90th percentile, or from 19.9 to 25.4	0.326	0.048	1.386	1.260	1.523
Course year level	From year 1 to year 4	-1.374	0.087	0.253	0.213	0.300
Percent grade	From 10th to 90th percentile (57-86%)	0.500	0.039	1.649	1.527	1.780
Mean of grade	From 10th to 90th percentile (57-86%)	2.031	0.086	7.625	6.444	9.023
Class size (ln(enrollment)	From 10th to 90th percentile (3.18-5.50)	-0.649	0.034	0.523	0.489	0.558
Mean Class size	From 10th to 90th percentile (3.18-5.50)	-0.327	0.102	0.721	0.590	0.881

Table 8. Random effects models for SET completion data – course type

Variable	**Co-efficient**	**Std. err.**	**Odds ratio**	**95% confidence interval**	
Fixed part					
Student age	0.059***	0.009	1.061	1.043	1.080
Course year level	-0.458***	0.029	0.633	0.598	0.670
Gender (male=1)	-0.545***	0.056	0.580	0.519	0.647
Term 1 or 2	-0.906***	0.023	0.404	0.386	0.423
Mean of Term by student	0.381***	0.076	1.463	1.261	1.698
Two-term course	-1.680***	0.052	0.186	0.168	0.206
Percent grade	0.017***	0.001	1.018	1.015	1.020
Mean of grade by student	0.070***	0.003	1.072	1.066	1.079
Class size (ln(enrollment))	-0.287***	0.015	0.751	0.729	0.773
Mean of class size	-0.144***	0.044	0.866	0.794	0.944
Individual study course	-0.649*	0.336	0.523	0.270	1.011
Experiential course	-0.362***	0.114	0.696	0.557	0.870
Small group course	0.004	0.064	1.004	0.886	1.137
Constant	-3.487***	0.367			
Random part			*Other statistics*		
$\sqrt{\hat{\psi}}$	3.295		N	94158	
ρ	0.767		Groups	21533	
Log likelihood	-47350.4		Obs/group	4.4	

Notes: Base category is Lecture course type.